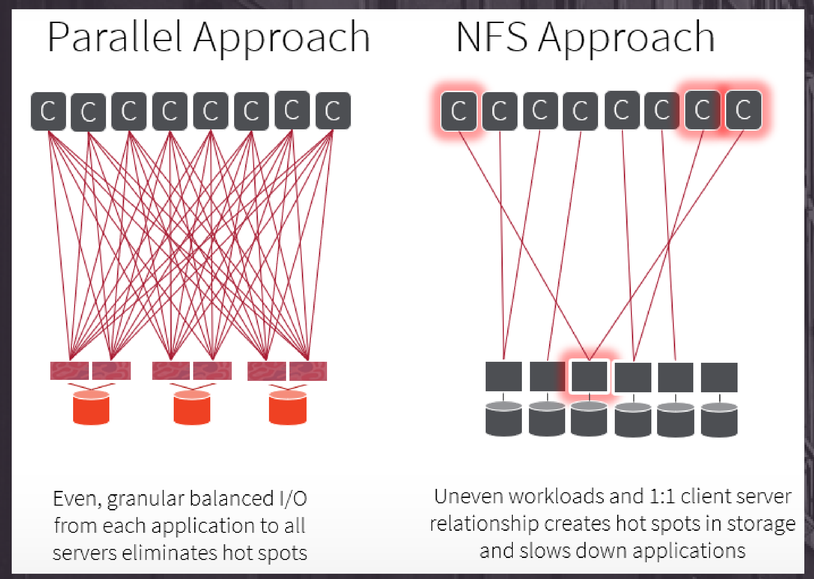
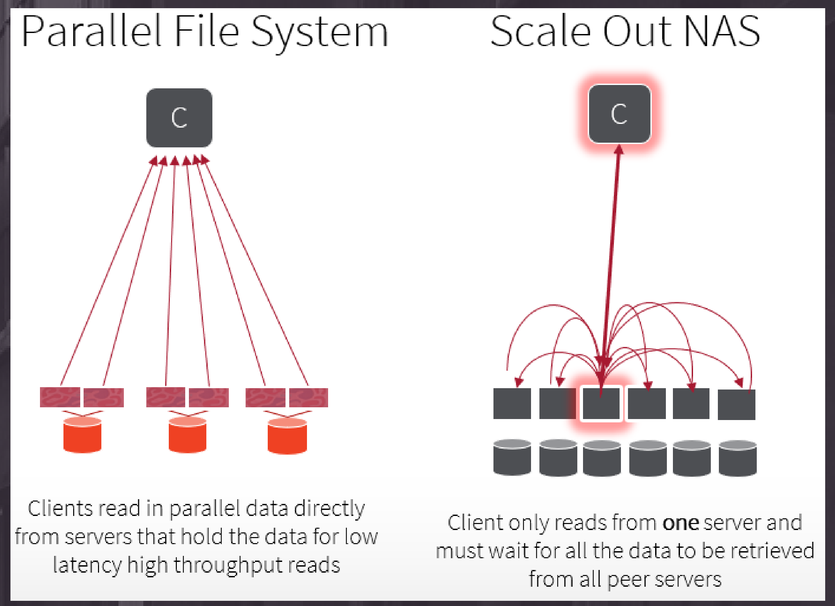
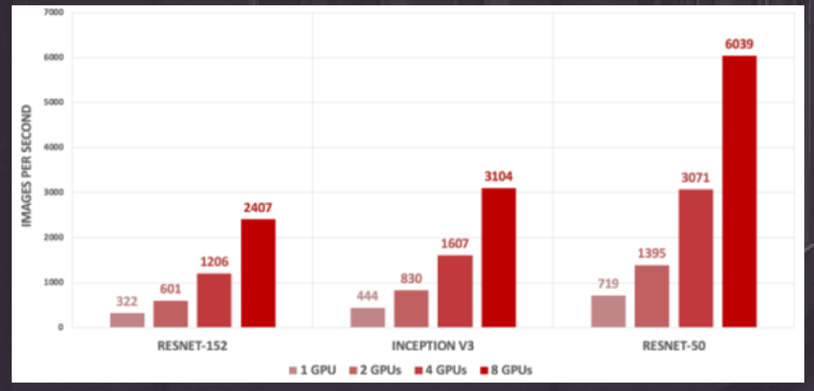
GPU Saturation Testing with Variable Applications and Storage Platforms

Poster Number:34 Authors: Brian Cox BRCox@DDN.com and Aaron Knister AKnister@DDN.com



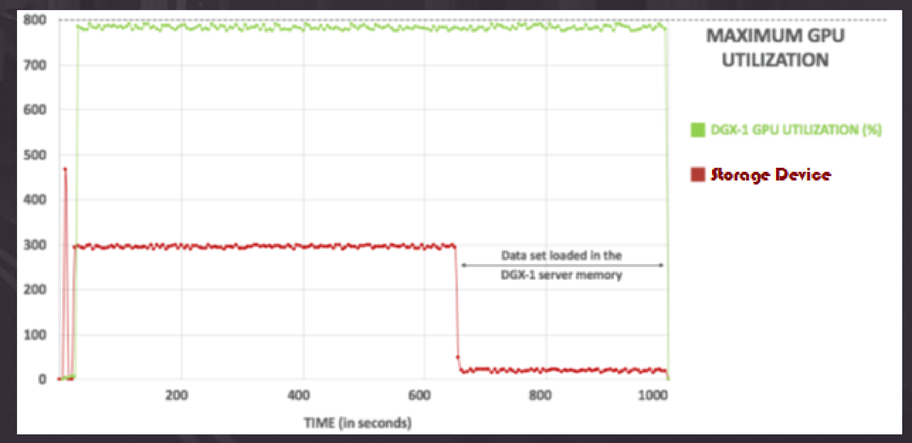
By design, the GPU architecture provides facilities for massive processing concurrency. Some GPU based applications distribute over 96 nodes simultaneously, touching 1500 GPUs. These application profiles necessitate parallel data paths that deliver data with high-throughput, low-latency and massive concurrency, directly to GPU memory.

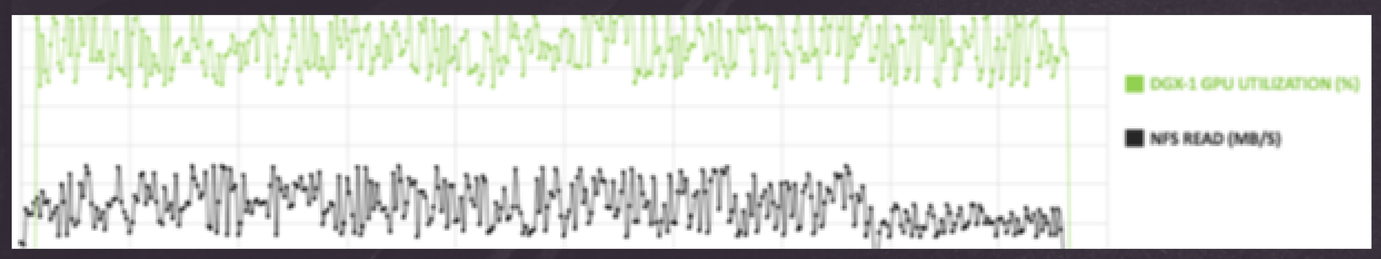




The graph on the left, demonstrates training application performance with resnet-50, resnet-152 and inceptionV3 models using different numbers of GPUs on a single DGX- 1 server. The resnet-152 and inceptionV3 tests were executed with the NVIDIA TensorFlow 18.03-py2 dockerfile and a data set from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The resnet-50 test was executed with the NVIDIA TensorFlow 18.09-py3 dockerfile and same data set

Illustrated to the right, is the GPU utilization and read activity from the ai-configured storage appliance. The GPUs achieve maximum utilization by using the storage appliance to deliver a steady stream of data through the training process. The application takes 933 seconds to complete. At approximately 660 seconds, the data set is fully loaded into the DGX-1 server and the application no longer needs to read the data from the storage appliance.





Illustrated to the left is the GPU utilization and read activity from the same application accessing data on NFS storage. The GPUs never achieve maximum utilization, and the NFS storage fails to deliver a steady stream of data to the application. The training application takes nearly 2000 seconds to complete.